Regresión lineal: enfoque Bayesiano

Matías Carrasco

14 de septiembre de 2023

Índice

1. La distribución gaussiana multivariada

Un vector de D atributos $\boldsymbol{x} = \left[x^{(1)}, \ldots, x^{(D)}\right]^\top$ tiene distribución gaussiana de **media $\boldsymbol{\mu}$** y **matriz de covarianzas $\boldsymbol{\Sigma}$** si su densidad en $\mathbb{R}^D$ está dada por la siguiente fórmula:

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

La media no es otra cosa que el vector de medias $\boldsymbol{\mu} = \left[\mu^{(1)}, \ldots, \mu^{(D)}\right]^\top$ de cada atributo y la matriz de covarianzas está dada por

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

y es por lo tanto una matriz simétrica y positiva.

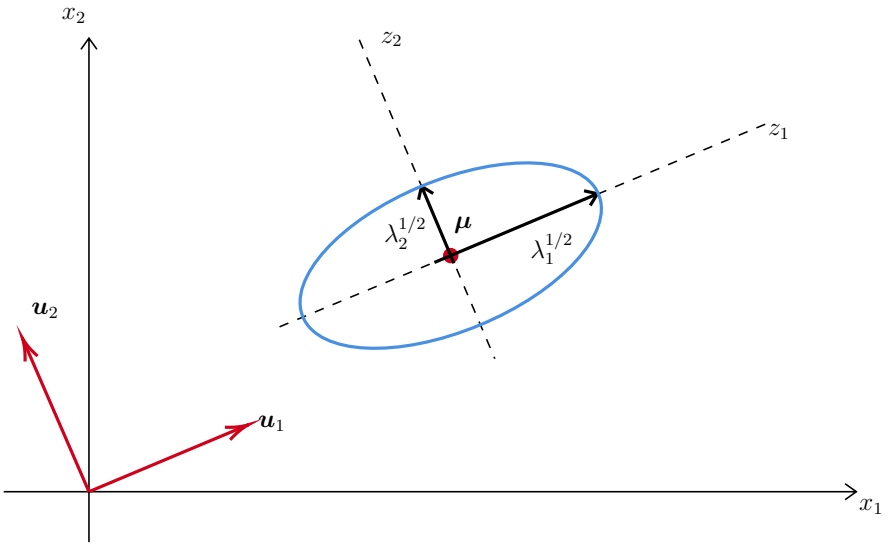

Figura 1: Curvas de nivel de la distribución gaussiana multivariada.

Existe una matriz ortogonal $\boldsymbol{U} = \left[\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(D)}\right]$ y una matriz diagonal $\boldsymbol{D} = [\lambda^{(1)}, \ldots, \lambda^{(D)}]$ tales que

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top = \sum_{j=1}^{D} \lambda^{(j)} \boldsymbol{u}^{(j)} \left(\boldsymbol{u}^{(j)}\right)^\top$$

Con esta notación es fácil ver que la inversa es:

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^\top = \sum_{j=1}^{D} \frac{1}{\lambda^{(j)}} \boldsymbol{u}^{(j)} \left(\boldsymbol{u}^{(j)}\right)^\top$$

Considerando las nuevas coordenadas

$$z^{(j)} = \left(\boldsymbol{u}^{(j)}\right)^\top (\boldsymbol{x} - \boldsymbol{\mu})$$

vemos que la expresión dentro de la exponencial toma la forma

$$-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{j=1}^{D} \frac{\left(z^{(j)}\right)^2}{\lambda^{(j)}}$$

De aquí vemos fácilmente que las curvas de nivel de la densidad son **elipses**, ver Fig. 1.

2. Regresión lineal bayesiana

2.1. El modelo

Consideremos el vector (columna) de atributos $\boldsymbol{x} = \left[1, x^{(1)}, \ldots, x^{(D)}\right]^\top$ y el vector (columna) de coeficientes $\boldsymbol{\theta} = [b, w_1, \ldots, w_D]^\top$, de modo tal que nuestro modelo lineal se escribe

$$y = b + w_1 x^{(1)} + w_2 x^{(2)} + \cdots + w_D x^{(D)} + \epsilon,$$

en donde $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Recordar que en notación matricial esto es

$$y = \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon$$

Apilando todas las observaciones podemos calcular todas las predicciones:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

en donde recordar que

$$\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\top$$

es la matroz de diseño;

$$\boldsymbol{y} = [y_1, \ldots, y_N]^\top$$

es el vector de targets o etiquetas, y

$$\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$$

el vector de ruidos.

Vamos a sumir (solo para simplificar la exposición) que σ^2 es un parámetro en el sentido clásico, y que además conocemos su valor.

La regla Bayes:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})\, p(\boldsymbol{\theta})}{p(\boldsymbol{y} \mid \boldsymbol{X})}$$

Si expandimos la verosimilitud marginal:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

Estamos interesados en hacer predicciones, por lo que para un vector de atributos $\boldsymbol{x}_{\text{nuevo}}$ la densidad de y_{nuevo} está dada por:

$$p(y_{\text{nuevo}} \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{X}, \boldsymbol{y}) = \int p(y_{\text{nuevo}} \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})\, d\boldsymbol{\theta}$$

Una vez calculada la densidad podemos dar predicciones:

$$P(a < y_{\text{nuevo}} \leq b \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{X}, \boldsymbol{y}) = \int_a^b p(z \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{X}, \boldsymbol{y})\, dz$$

2.2. Verosimilitud

La verosimilitud $p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})$ viene dada por nuestro modelo

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

donde $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$. Es decir:

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X}) = \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_N\right),$$

una Gaussiana multivariada en dimensión N, con media $\boldsymbol{X}\boldsymbol{\theta}$ y matriz de covarianzas $\sigma^2 \mathbf{I}_N$.

2.3. A priori

Vamos a elegir una distribución a priori $p(\boldsymbol{\theta})$ que sea conjugada a la verosimilitud:

$$p(\boldsymbol{\theta}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),$$

cuyos parámetros son la media $\boldsymbol{\mu}_0$ y la matriz de covarianzas $\boldsymbol{\Sigma}_0$.

2.4. A posteriori

Vamos a calcular la distribución a posteriori:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto & p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})\, p(\boldsymbol{\theta}) \\ = & \frac{1}{(2\pi)^{N/2} \left|\sigma^2 \mathbf{I}\right|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top \left(\sigma^2 \mathbf{I}\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\right) \\ & \times \\ & \frac{1}{(2\pi)^{N/2} \left|\boldsymbol{\Sigma}_0\right|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)\right) \end{aligned}$$

Eliminando las constantes nos quedamos con:

$$\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\right) \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)\right)$$

O de forma equivalente:

$$\exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)\right)\right\}$$

Desarrollando y descartando todo lo que no tenga $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\theta} + \frac{1}{\sigma^2}\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta} - 2\boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta}\right)\right\}$$

Sabemos que la distribución a posteriori es Gaussiana, digamos de parámetros $\boldsymbol{\mu}_1$ y $\boldsymbol{\Sigma}_1$:

$$\begin{aligned} p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}\right) &= \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\mu}_1\right)^\top \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\theta}-\boldsymbol{\mu}_1\right)\right) \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\theta} - 2\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\theta}\right)\right\} \end{aligned}$$

Igualando los términos cuadráticos podemos despejar $\boldsymbol{\Sigma}_1$:

$$\begin{aligned} \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\theta} &= \frac{1}{\sigma^2}\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta} \\ &= \boldsymbol{\theta}^\top\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\theta} \end{aligned}$$

de donde obtenemos:

$$\boxed{\boldsymbol{\Sigma}_1 = \left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}}$$

De forma similar, igualando los términos lineales y usando la expresión para $\boldsymbol{\Sigma}_1$ obtenemos una expresión para $\boldsymbol{\mu}_1$:

$$\begin{aligned} -2\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\theta} &= -\frac{2}{\sigma^2}\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\theta} - 2\boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta} \\ \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} &= \frac{1}{\sigma^2}\boldsymbol{y}^\top \boldsymbol{X} + \boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}_0^{-1} \\ \boldsymbol{\mu}_1^\top &= \left(\frac{1}{\sigma^2}\boldsymbol{y}^\top \boldsymbol{X} + \boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_1 \end{aligned}$$

de donde obtenemos

$$\boxed{\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)}$$

En resumen:

$$\boxed{p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}\right) = \mathcal{N}\left(\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right), \left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\right)}$$

2.5. Comparación con la regularización Ridge

Si elegimos $\boldsymbol{\mu}_0 = \mathbf{0}$ y $\boldsymbol{\Sigma}_0 = \frac{\sigma^2}{N\alpha}\mathbf{I}$ obtenemos que la media de la distribución a posteriori es

$$\boldsymbol{\mu}_1 = \left(\boldsymbol{X}^\top \boldsymbol{X} + N\alpha\mathbf{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

que es igual a la solución del problema de regresión de Ridge:

$$\widehat{\boldsymbol{\theta}}_\alpha = \begin{bmatrix} \widehat{b} \\ \widehat{\boldsymbol{w}} \end{bmatrix} = \arg\min_{\boldsymbol{\theta}} \Big\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\| + \alpha\|\boldsymbol{w}\|^2 \Big\}.$$

Es decir, la regresión bayesiana ya tiene incluida una forma de regularización con la consideración de una distribución a priori.

2.6. Predicciones

Dada una observación nueva $\boldsymbol{x}_{\text{nuevo}}$, nos interesa la densidad:

$$p\left(y_{\text{nuevo}} \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{X}, \boldsymbol{y}\right).$$

Una forma de hacerlo es integrar respecto de $\boldsymbol{\theta}$ ponderando por la distribución a posteriori $p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}\right)$:

$$\begin{aligned} p\left(y_{\text{nuevo}} \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{X}, \boldsymbol{y}\right) &= \mathbb{E}_{\text{A posteriori}}\left\{p\left(y_{\text{nuevo}} \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{\theta}\right)\right\} \\ &= \int p\left(y_{\text{nuevo}} \mid \boldsymbol{x}_{\text{nuevo}}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}\right) d\boldsymbol{\theta} \end{aligned}$$

Sin embargo, nuestro modelo es:

$$y_{\text{nuevo}} = \boldsymbol{x}_{\text{nuevo}}^\top \boldsymbol{\theta} + \epsilon$$

con $\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$ y $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. Además estos términos son independientes.

Cómo la proyección $\boldsymbol{x}^\top \boldsymbol{\theta}$ es normal, la distribución de y_{nuevo} es también normal.

Su media es:

$$\mathbb{E}_{\text{A posteriori}}\left[y_{\text{nuevo}}\right] = \mathbb{E}_{\text{A posteriori}}\left[\boldsymbol{x}_{\text{nuevo}}^\top \boldsymbol{\theta}\right] = \boldsymbol{x}_{\text{nuevo}}^\top \mathbb{E}_{\text{A posteriori}}\left[\boldsymbol{\theta}\right] = \boldsymbol{x}_{\text{nuevo}}^\top \boldsymbol{\mu}_1$$

Su varianza es:

$$\mathbb{E}_{\text{A posteriori}}\left[y_{\text{nuevo}}^2\right] - \left(\boldsymbol{x}_{\text{nuevo}}^\top \boldsymbol{\mu}_1\right)^2 = \mathbb{E}_{\text{A posteriori}}\left[\left(\boldsymbol{x}_{\text{nuevo}}^\top \boldsymbol{\theta} + \epsilon\right)^2\right] - \left(\boldsymbol{x}_{\text{nuevo}}^\top \boldsymbol{\mu}_1\right)^2$$

El primer término del lado derecho es igual a

$$\begin{aligned}\mathbb{E}_{\text{A posteriori}}\left[\left(\boldsymbol{x}_{\text{nuevo}}^{\top}\boldsymbol{\theta}\right)^2\right]+\sigma^2 &= \mathbb{E}_{\text{A posteriori}}\left[\boldsymbol{x}_{\text{nuevo}}^{\top}\boldsymbol{\theta}\boldsymbol{\theta}^{\top}\boldsymbol{x}_{\text{nuevo}}\right]+\sigma^2\\ &= \boldsymbol{x}_{\text{nuevo}}^{\top}\,\mathbb{E}_{\text{A posteriori}}\left[\boldsymbol{\theta}\boldsymbol{\theta}^{\top}\right]\boldsymbol{x}_{\text{nuevo}}+\sigma^2\\ &= \boldsymbol{x}_{\text{nuevo}}^{\top}\left(\boldsymbol{\Sigma}_1+\boldsymbol{\mu}_1\boldsymbol{\mu}_1^{\top}\right)\boldsymbol{x}_{\text{nuevo}}+\sigma^2\\ &= \boldsymbol{x}_{\text{nuevo}}^{\top}\boldsymbol{\Sigma}_1\boldsymbol{x}_{\text{nuevo}}+\left(\boldsymbol{x}_{\text{nuevo}}^{\top}\boldsymbol{\mu}_1\right)^2+\sigma^2\end{aligned}$$

y finalmente obtenemos que la varianza es:

$$\boldsymbol{x}_{\text{nuevo}}^{\top}\boldsymbol{\Sigma}_1\boldsymbol{x}_{\text{nuevo}}+\sigma^2$$

En resumen

$$\boxed{p\left(y_{\text{nuevo}}\mid\boldsymbol{x}_{\text{nuevo}},\boldsymbol{X},\boldsymbol{y}\right)=\mathcal{N}\left(\boldsymbol{x}_{\text{nuevo}}^{\top}\boldsymbol{\mu}_1,\boldsymbol{x}_{\text{nuevo}}^{\top}\boldsymbol{\Sigma}_1\boldsymbol{x}_{\text{nuevo}}+\sigma^2\right)}$$

2.7. Verosimilitud marginal

La verosimilitud marginal puede utilizarse para seleccionar hiperparámetros. A modo de ejemplo comentaremos sobre su utilización en la selección del grado en una regresión polinomial.

Para un polinomio de grado K el modelo de ruido Gaussiano es:

$$y=b+w_1x_n+w_2x^2+\cdots+w_Kx^K+\epsilon$$

donde $\epsilon\sim\mathcal{N}\left(0,\sigma^2\right)$.

En forma vectorial:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\theta}+\boldsymbol{\epsilon},$$

donde $\boldsymbol{\theta}=\left[b,w_1,\ldots,w_K\right]^{\top}$, $\boldsymbol{x}=\left[1,x,x^2,\ldots,x^K\right]^{\top}$, $\boldsymbol{y}=\left[y_1,\ldots,y_N\right]^{\top}$ y la matriz de diseño, $\boldsymbol{X}=\left[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\right]^{\top}$.

La verosimilitud marginal está dada por la integral

$$p\left(\boldsymbol{y}\mid\boldsymbol{X}\right)=\int p\left(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}.$$

De forma análoga a las cuentas que hemos hecho hasta ahora se puede ver que:

$$\boxed{p\left(\boldsymbol{y}\mid\boldsymbol{X}\right)=\mathcal{N}\left(\boldsymbol{X}\boldsymbol{\mu}_0,\sigma^2\mathbf{I}_N+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^{\top}\right)}$$

Luego podemos seleccionar el modelo con mayor verosimilitud marginal.